Extractive Multi-Document Summarization with Integer Linear Programming and Support Vector Regression

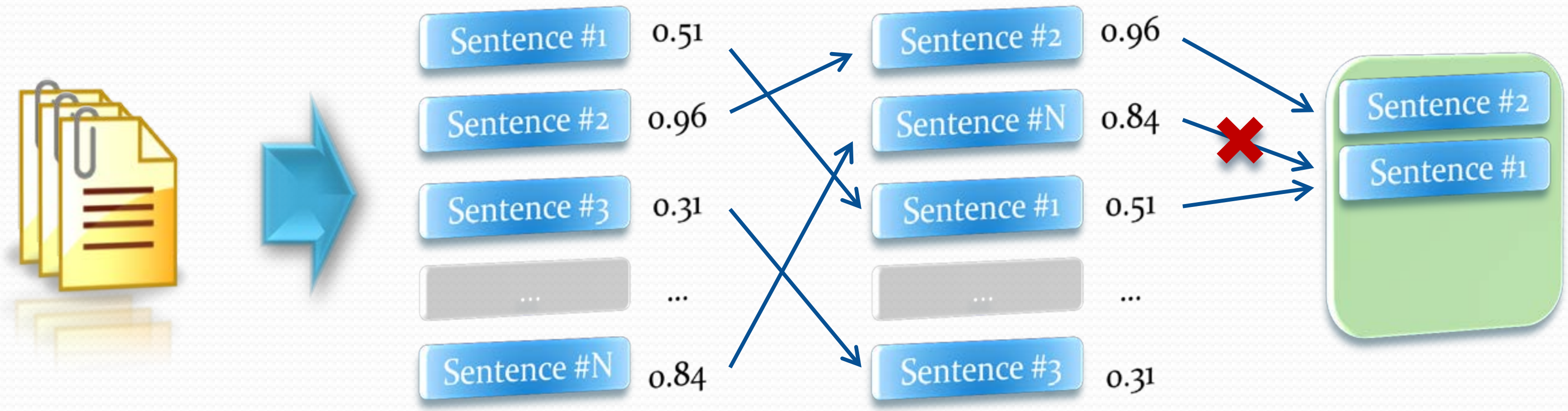
Dimitrios Galanis, Gerasimos Lampouras and Ion Androutsopoulos

Multi-document Summarization

- We aim to produce **summaries** that are:
  - **relevant** to the query,
  - **diverse** (do not repeat information),
  - **grammatical**,
  - and up to a certain **length**.
- An **extractive** summarization system
  - includes only **un-altered sentences**.
- An **abstractive** summarization system
  - **may alter** (shorten, paraphrase, etc.) sentences,
  - requires **more processing time**,
  - usually requires **specialized resources** (parsers, paraphrasing rules etc.),
  - is in practice, **marginally better** than an extractive system.

Greedy Approach to Summarization

- Many extractive summarization systems use a **greedy approach**.
  - They maximize the **importance** of the summary's **sentences**.
  - Importance can be estimated via **statistics**, **machine learning** etc.



- **Sentence diversity** can be achieved by discarding any sentence that is **too similar** to the sentences already in the summary.
  - **Similarity measures** (e.g., cosine similarity) are often employed.

- We use the **greedy approach** as a **baseline**.
  - We present a <u>**non-greedy**</u> approach, based on **global optimization**.

Global Optimization Approach

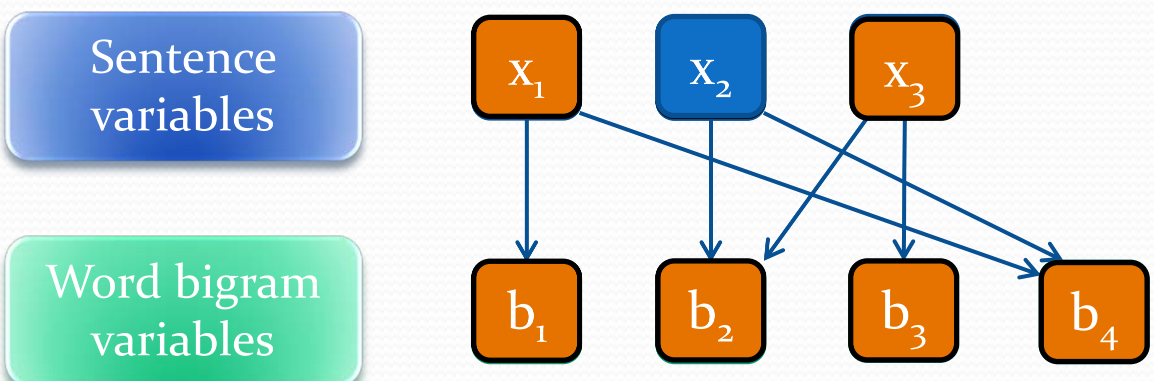
- Recent work shows that **global optimization** approaches produce **better** (or comparable) summaries, compared to greedy approaches.
  - Take into account the **entire search space** to find an **optimal** solution.

- We jointly optimize **sentence importance** and **diversity** to find an optimal summary.
  - Respecting the **maximum summary length**.

- We do **extractive summarization**, we do not alter the source sentences.
  - But optimization models can be **easily extended.**
  - Sentence **compression**, sentence **aggregation** etc.

ILP-Based Global Optimization

- We use **Integer Linear Programming** (**ILP**).
  - **Binary LP**: all the **variables** are **binary (0/1)**.
- We **maximize** the summary's **Imp(S) + Div(S)**.
  - **Imp(S)**: Sum of importance scores of sentences in summary S.
  - **Div(S)**: Sum of **distinct** selected **word bigrams** in summary S.
    - Following previous work, we assume that **bigrams** roughly **correspond** to **concepts/things**.



Sentence	Importance
x_1	0.8
x_2	0.7
x_3	0.6

~~Importance~~	~~1.5~~
~~Diversity~~	~~3~~

Importance	1.4
Diversity	4

ILP Objective function

$\lambda_1 + \lambda_2 = 1$

$$\max \lambda_1 \cdot imp(S) + \lambda_2 \cdot div(S) =$$

Sentence variable (0/1)

Bigram variable (0/1)

$$\max_{b,x} \lambda_1 \cdot \sum_{i=1}^{n} a_i \cdot \frac{l_i}{L_{\max}} \cdot x_i + \lambda_2 \cdot \sum_{i=1}^{|B|} \frac{b_i}{n}$$

Sentence importance score, ranges in [0, 1].

Normalized sentence length: Rewards **longer** sentences.

Number of input sentences

ILP Constraints

$$\text{subject to} \sum_{i=1}^{n} l_i \cdot x_i \leq L_{max}$$

The summary length **must not exceed** the **maximum allowed length.**

$$\text{and} \sum_{g_j \in B_i} b_j \geq |B_i| \cdot x_i, \text{ for } i = 1 \ldots n$$

$$\text{and} \sum_{s_i \in S_j} x_i \geq b_j, \text{ for } j = 1 \ldots |B|$$

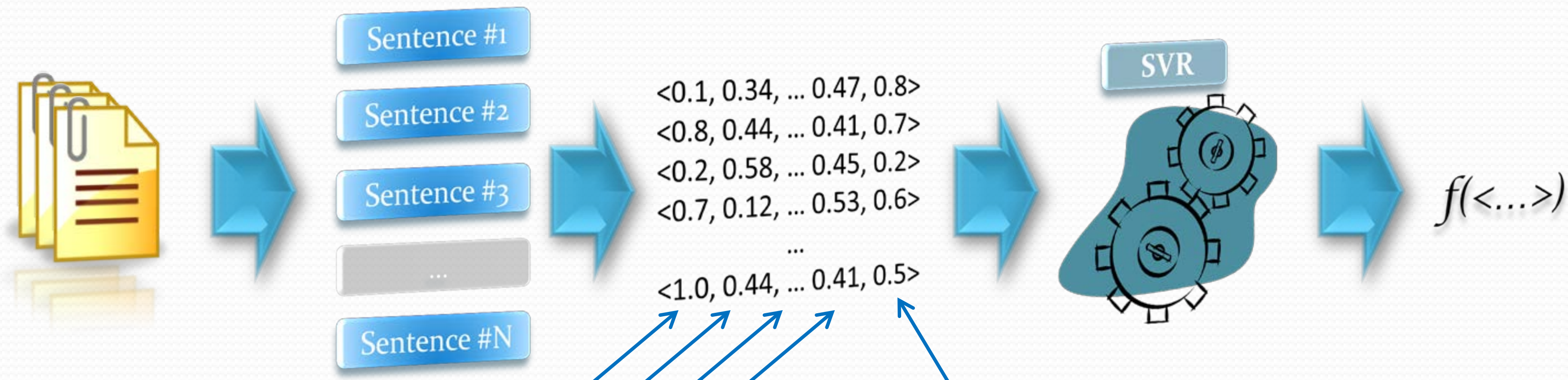
Constrains to ensure **consistency** between **sentences** and **bigrams**.

If a **sentence** is **included,** **all the bigrams** it contains must also be **included**.

If a **bigram** is included, at **least one sentence** that contains it must also be **included**.

SVR Model of Sentence Importance

- **SVR – Support Vector Regression**
  - Regression equivalent of **Support Vector Machines**.
  - Rather than **classification**, it aims to learn a **function** with **real values**.



Feature vector, one per candidate **sentence**:
- **sentence position** in the original document,
- number of **named entities**,
- **Levenshtein** distance between **query** and **sentence**,
- **word overlap** between **query** and **sentence**,
- content **word** and **document frequencies**.

Target sentence importance score:
- The **SVR** learns to **predict** this value.
- Similarity between **sentence** and **human-written** summaries.

Estimated as the average of **ROUGE-2** and **ROUGE-SU4** scores.
- **Bigram** similarity measures.
- **Highly correlated** with human judgments in **summarization**.

Evaluation Setup

- We **experimented** with the following **systems** and **baselines**:
  - **ILP** system.
  - **GREEDY** system.
    - Uses the same **SVR (for importance scores)** as the ILP system.
  - **GREEDY-RED** system.
    - Includes **redundancy checks** via cosine similarity.

- Datasets: **DUC 2005**, **DUC 2006**, **DUC 2007** and **TAC 2008**.
  - Each dataset contains **queries** and corresponding **sets of relevant documents**.
  - For each **query**, multiple **reference (human-authored) summaries** are also provided.

Efficiency

- **Our ILP method** is a generalization of **0-1 Knapsack** (**NP-Hard**).
  - But we input only the **top 100 sentences** with the **highest SVR scores**.
  - We also **ignore** in the ILP model **bigrams** that consist exclusively of **stop words** or occur only **once**.
  - The steps above **reduce** the ILP variables to the **order of hundreds**.
  - The ILP variables grow **approximately linearly** to the **number** and **length** of the input sentences.

- **0.9 - 1.25 seconds** are required for an **off-the-shelf solver** to find the optimal solution **per summary**.
  - If we include **preprocessing** of input documents and **formulation** of the **ILP program**, it takes **10-11 seconds** to produce a summary.

Results on the Development Set

- In all cases, we **trained the SVR** on **DUC 2006** data.
- We used **DUC 2007** as a **development set** for parameter tuning.
  - Best results are achieved for $\lambda_1 = 0.4, \lambda_2 = 0.6$.
  - Both **sentence importance** and **diversity** contribute to the results.

system	ROUGE-2	ROUGE-SU4
ILP ($\lambda_1 = 0.4$)	**0.12517**	0.17603
GREEDY-RED	0.11591	0.16908
GREEDY	0.11408	0.16651
Lin and Bilmes 2011	0.12380	N/A
Celikyilmaz and Hakkani-Tur 2010	0.11400	0.17200
Haghighi and Vanderwende 2009	0.11800	0.16700
Schilder and Ravikumar 2008	0.11000	N/A
Pingali et al. 2007 (DUC 2007)	0.12448	**0.17711**
Toutanova et al. 2007 (DUC 2007)	0.12028	0.17074
Conroy et al. 2007 (DUC 2007)	0.11793	0.17593
Amini and Usunier 2007 (DUC 2007)	0.11887	0.16999

Our **ILP method outperforms** the **baselines**.

Our ILP method has the **best ROUGE-2**.

And the **second best ROUGE-SU4** score.

But these are **development set results**.

Results on Test Set – TAC 2008

system	ROUGE-2	ROUGE-SU4
ILP ($\lambda_1 = 0.4$)	0.11168	0.14413
Woodsend and Lapata 2012 (with QSTG)	0.11370	**0.14470**
Woodsend and Lapata 2012 (without QSTG)	0.10320	0.13680
Berg-Kirkpatrick et al. 2011 (with subtree cuts)	**0.11700**	0.14380
Berg-Kirkpatrick et al. 2011 (without subtree cuts)	0.11050	0.13860
Shen and Li 2010	0.09012	0.12094
Gillick and Favre 2009 (with sentence compression)	0.11100	N/A
Gillick and Favre 2009 (without sentence compr.)	0.11000	N/A
Gillick et al. 2008 (run 43 in TAC 2008)	0.11140-	0.14298-
Gillick et al. 2008 (run 13 in TAC 2008)	0.11044-	0.13985-
Conroy and Schlesinger 2008 (run 60 in TAC 2008)	0.10379-	0.14200-
Conroy and Schlesinger 2008 (run 37 in TAC 2008)	0.10338-	0.14277-
Conroy and Schlesinger 2008 (run 06 in TAC 2008)	0.10133+	0.13977-
Galanis and Malakasiotis 2008 (run 02 in TAC 2008)	0.10012+	0.13694-

Third best results in **ROUGE-2** and **second best** in **ROUGE-SU4**.

Some methods are **abstractive**.

Best results amongst **extractive**.

Better results than some **abstractive**.

+ and - denote the existence or absence of statistical significance (t-test), respectively.

Conclusions

- We presented an **ILP-based** method for **multi-document extractive summarization** that **jointly maximizes**:
  - **sentence importance** scores provided by a **Support Vector Regression (SVR)** model, and
  - **sentence diversity** scores, computed as the number of **distinct bigrams of the input documents** that occur in the summary,
  - respecting the **maximum allowed summary length**.
- **Experiments** on widely used **benchmark datasets** show that our **ILP-based method:**
  - achieves **state of the art results** amongst **extractive** methods,
  - **outperforms** two **greedy baselines** that use the **same SVR model** (without ILP),
  - **performs better** than some **abstractive** methods.
- **Future** work:
  - We are experimenting with an **extended form of our ILP-based method** that **includes sentence compression** (Galanis & Androutsopoulos 2010).

Thank you!

Questions?